Available online at www.sciencedirect.com

ScienceDirect

Journal of Multivariate Analysis 97 (2006) 1997–2008

Journal of Multivariate Analysis

www.elsevier.com/locate/jmva

Image classification based on Markov random field models with Jeffreys divergence

Ryuei Nishii[a,*], Shinto Eguchi[b]

[a] *Faculty of Mathematics, Kyushu University, Hakozaki, Higashiku Fukuoka 812-8581, Japan*
[b] *Institute of Statistical Mathematics, Minami-Azabu, Minatoku Tokyo 106-8569, Japan*

Received 21 April 2005
Available online 7 July 2006

Abstract

This paper considers image classification based on a Markov random field (MRF), where the random field proposed here adopts Jeffreys divergence between category-specific probability densities. The classification method based on the proposed MRF is shown to be an extension of Switzer's soothing method, which is applied in remote sensing and geospatial communities. Furthermore, the exact error rates due to the proposed and Switzer's methods are obtained under the simple setup, and several properties are derived. Our method is applied to a benchmark data set of image classification, and exhibits a good performance in comparison with conventional methods.

AMS 2000 subject classification: 62H30; 62H35

Keywords: Bayes estimate; Discriminant analysis; Image analysis; Kullback–Leibler information

1. Introduction

Consider a single image such that multivariate data (*feature vectors*) are observed at respective pixels (*multicolored images*). Image classification is a problem of classifying pixels into several homogeneous regions by learning the feature vectors and the adjacency relationships of the pixels in the image. The classification of a pixel into one of categories is an important and fundamental problem in image pattern analysis, see, e.g., [10].

In image classification, normal distributions are frequently used for analyzing multivariate data in a feature space, and Markov random fields (MRFs) are used for modeling the distribution of

* Corresponding author.
E-mail addresses: nishii@math.kyushu-u.ac.jp (R. Nishii), eguchi@ism.ac.jp (S. Eguchi).

0047-259X/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2006.01.009

categories in the image, see [9,11]. It is usually assumed that the feature vectors conditional on category labels are independent (*conditional independence*) and the labels follow the MRF. Thus, the joint probability density function is obtained by the product of two density functions. The estimates of parameters specifying category-specific distributions can be easily formulated due to the assumption. On the other hand, the estimation of parameters specifying the MRF is not an easy task because the probability distribution cannot be expressed in a closed form. Hence, the pseudo-likelihood is frequently used for this purpose.

The key issue is the estimation of pixel labels of test data. Computer-intensive methods including simulated annealing [5] and the iterative conditional mode (ICM) method [1] can be used for the estimation, but the implementation is often difficult because of computational complexity. See [2,3] for spatial statistics, and [4,7,12,14] for discriminant analysis.

The Jeffreys divergence [8] gives the inherited distance between category-specific distributions. The divergence is incorporated into MRF modeling. Our modeling introduces a parsimonious expression for the distances among the categories to reduce the model complexity. Our image classification method will be shown to give a new insight into Switzer's smoothing method [15]. Furthermore, our method is applied to a benchmark data set for classification, and it exhibits an efficient performance in comparison with conventional methods.

This paper is organized as follows. Section 2 reviews basic distributional assumptions in image classification and introduces MRFs specified by the divergences between category-specific densities. The relationship between the proposed classification method and Switzer's smoothing method is discussed in Section 3. Exact error rates of classification are obtained under the simple setup. The use of spatial information is proved to always improve noncontextual classification, even if the estimate of the spatial-dependency parameter is far from the true value. Parameter estimation procedures are given in Section 4. The proposed classification method is applied to a benchmark data set in Section 5, and it exhibits better performance than that of the Potts model. Finally, the paper is concluded in Section 6. Derivations of exact error rates and parameter estimation methods are given in the Appendix.

2. MRFs based on Jeffreys divergence

The objective of this section is to introduce MRFs based on divergence (*divergence models*) for contextual image classification.

2.1. Basic distributional assumptions imposed on spatial data

Let $\mathcal{D}$ be a training area consisting of n pixels. The training pixels are numbered from 1 to n, and set $\mathcal{D} = \{1, \ldots, n\}$. Each pixel i in area $\mathcal{D}$ is supposed to belong to one of G categories $C_1, \ldots, C_G$. Suppose that an m-dimensional multivariate feature vector $\boldsymbol{x}_i \in \mathbb{R}^m$ is observed at each pixel i. A label of the category covering pixel i is denoted by y_i in the label set $\mathcal{G} \equiv \{1, \ldots, G\}$. Thus, the set $\{(\boldsymbol{x}_i, y_i) \in \mathbb{R}^m \times \mathcal{G} | i \in \mathcal{D}\}$ constitutes a training data set. An image with $n = 9$ pixels is given in Fig. 1, and category labels in the same image are given in Fig. 2.

Random vectors and variables corresponding to $\boldsymbol{x}_i$ and y_i are, respectively, denoted by $\boldsymbol{X}_i$ and Y_i. Set $\boldsymbol{X} = (\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_n^T)^T$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$. The joint distribution $p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x}, \boldsymbol{y})$ of the feature vector $\boldsymbol{X}$ and the label vector $\boldsymbol{Y}$ can then be decomposed to $p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x}, \boldsymbol{y}) = p_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x} \mid \boldsymbol{y}) \cdot p_{\boldsymbol{Y}}(\boldsymbol{y})$. The fundamental assumption is imposed on the conditional independence of $\boldsymbol{X}$ given $\boldsymbol{Y} = \boldsymbol{y}$. That is,

$$p_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y}) = \prod_{i \in \mathcal{D}} f(\boldsymbol{x}_i, \boldsymbol{\theta}(y_i)), \tag{1}$$

5	4	3
6	1	2
7	8	9

Fig. 1. Pixel numbers of the image of size $n = 9$.

1	1	1
1	1	2
1	2	2

Fig. 2. Pixel labels of the image Fig. 1.

where $f(\boldsymbol{x}, \boldsymbol{\theta}(g))$ is a *category-specific probability density function*, and $\boldsymbol{\theta}(g)$ is an unknown parameter vector.

Next, label vector $\boldsymbol{Y}$ is assumed to follow a pairwise-dependent MRF with a neighborhood system $\{\mathcal{N}_i \subset \mathcal{D} \mid i \in \mathcal{D}\}$, where $\mathcal{N}_i$ denotes a neighborhood of pixel i. Let $J(g, h)$ be a quasi-distance between the categories C_g and C_h, and $\boldsymbol{Y}_{-i} : (n-1) \times 1$ be a vector $\boldsymbol{Y}$ of all the labels except Y_i. Then, we assume that the distribution of label Y_i conditional on $\boldsymbol{Y}_{-i} = \boldsymbol{y}_{-i}$ is specified by labels in neighborhood $\mathcal{N}_i$ of pixel i as

$$\Pr\{Y_i = g \mid \boldsymbol{Y}_{-i} = \boldsymbol{y}_{-i}\} = \frac{\exp\{-\beta\Delta_i(g)\}}{\sum_{g' \in \mathcal{G}} \exp\{-\beta\Delta_i(g')\}}, \quad \text{set } p_i(g \mid \boldsymbol{y}_{-i}), \tag{2}$$

where β is a non-negative constant called *a clustering parameter* or a granularity, and $\Delta_i(g)$ denotes the average of pseudo-distances between the center pixel with label g and neighbors in $\mathcal{N}_i$ as follows:

$$\Delta_i(g) = \sum_{h \in \mathcal{G}} r_i(h) J(h, g), \quad r_i(h) = |\{j \in \mathcal{N}_i \mid y_j = h\}|/|\mathcal{N}_i|. \tag{3}$$

Note that constant $r_i(h)$ denotes a relative frequency of pixels with label h in the neighborhood $\mathcal{N}_i$. The conditional probability of formula (2) becomes large if the average $\Delta_i(g)$ is small, e.g., most of the neighbors are labeled by g, which is the label of the center pixel. β gives the degree of spatial dependency of the MRF. If $\beta = 0$, the categories are spatially independent.

In general, the Hammersley–Clifford theorem assures that the conditional distribution of formula (2) specifies the joint distribution of label vector $\boldsymbol{Y}$ under the mild condition. Labels of test data are estimated by the *maximum a posteriori* (MAP) principle. Simulated annealing described in [5] and the ICM method described in [1] are the principal estimation approaches.

2.2. The Potts model and Jeffreys divergence

The simplest quasi-distance, $J(g, h)$, is the 0–1 distance defined by $J_0(g, h) = 1 - \delta_{gh}$, where δ_{gh} stands for Kronecker's delta. The spatial model with the distance $J_0(g, h)$ is the Ising model, which is one of the pillars of statistical mechanics. The model, however, is not always suitable in a case having more than two categories (Potts model). Therefore, Nishii [13] proposed to take the quasi-distance by the squared Mahalanobis distance. This approach will be extended into a more general setting.

In this paper, Jeffreys divergence [8] between category-specific densities is proposed for use as a quasi-distance. The divergence between two probability densities is defined as follows:

$$J(g,h)=\int\{f(\boldsymbol{x},\boldsymbol{\theta}(g))-f(\boldsymbol{x},\boldsymbol{\theta}(h))\}\log\left\{\frac{f(\boldsymbol{x},\boldsymbol{\theta}(g))}{f(\boldsymbol{x},\boldsymbol{\theta}(h))}\right\}d\boldsymbol{x}\geqslant 0. \tag{4}$$

This is the symmetrized Kullback–Leibler divergence, and the equality holds if and only if $g=h$. See [16] for further properties.

As an example, assume that the category-specific distribution is given by a normal distribution with mean vector $\boldsymbol{\mu}(g)$ and common covariance matrix Σ denoted by $N_m(\boldsymbol{\mu}(g),\Sigma)$, which is *homoscedastic*. In the homoscedastic case, Jeffreys divergence is reduced to the squared Mahalanobis distance $D(\boldsymbol{s},\boldsymbol{t};\Sigma)=(\boldsymbol{s}-\boldsymbol{t})^T\Sigma^{-1}(\boldsymbol{s}-\boldsymbol{t})$. In the heteroscedastic case, the divergence between distributions $N_m(\boldsymbol{\mu}(g),\Sigma(g))$ and $N_m(\boldsymbol{\mu}(h),\Sigma(h))$ is given by formula B.6 in the Appendix.

3. Gaussian MRFs and error estimates

Refer to Fig. 1 again as an image with center pixel 1 and its neighbors. First- and second-order neighborhoods of the center are given by sets of pixel numbers $\{2,4,6,8\}$ and $\{2,3,\ldots,9\}$, respectively. We focus our attention on center pixel 1 and its neighborhood $\mathcal{N}_1$ of size $2K$. We discuss the classification problem of center pixel 1 when labels y_j of neighbors in $\mathcal{N}_1$ are observed.

3.1. Relationship between the divergence model and Switzer's model

Let us consider the divergence model in Gaussian MRFs (GMRFs) where feature vectors follow homoscedastic normal distributions $N_m(\boldsymbol{\mu}(g),\Sigma)$. The divergence model will be shown to be a natural extension of Switzer's model.

Let $\widehat{\beta}$ be a non-negative estimated value of clustering parameter β. Then, label y_1 of center pixel 1 is estimated by the ICM algorithm. In this case, the estimate is derived by maximizing the product of normal density $\psi(\boldsymbol{x}_1;\boldsymbol{\mu}(g),\Sigma)$ and conditional probability $p_1(g\mid\boldsymbol{y}_{-1})$ defined by formula (2). This is equivalent to finding label $\widehat{Y}_{\mathrm{Div}}$ defined by

$$\widehat{Y}_{\mathrm{Div}}=\arg\min_{g\in\mathcal{G}}\left\{D(\boldsymbol{x}_1,\boldsymbol{\mu}(g);\Sigma)+\frac{\widehat{\beta}}{K}\sum_{j\in\mathcal{N}_1}D(\boldsymbol{\mu}(y_j),\boldsymbol{\mu}(g);\Sigma)\right\}, \tag{5}$$

where $D(\boldsymbol{s},\boldsymbol{t};\Sigma)$ is the squared Mahalanobis distance.

Switzer's model [15] is based on the local continuity assumption: "any center pixel and its neighbors jointly belong to the same category." See Fig. 1 as an example. If center pixel 1 has label g, then pixel labels in the first-order neighborhood are given by $y_2=y_4=y_6=y_8=g$. Furthermore, feature vectors observed at the center pixel and the neighbors are assumed to be independent. Then, the center pixel is classified by the majority vote of likelihoods by maximizing $\log\psi(\boldsymbol{x}_1;\boldsymbol{\mu}(g),\Sigma)+\sum_{j\in\mathcal{N}_1}\log\psi(\boldsymbol{x}_j;\boldsymbol{\mu}(g),\Sigma)$ with respect to label g. Here, Switzer's model can be slightly extended by changing the coefficient for $\sum_{j\in\mathcal{N}_1}\log\psi(\boldsymbol{x}_j;\boldsymbol{\mu}(g),\Sigma)$ from one to $\widehat{\beta}/K$. Thus, we define:

$$\widehat{Y}_{\mathrm{Switzer}}=\arg\min_{g\in\mathcal{G}}\left\{D(\boldsymbol{x}_1,\boldsymbol{\mu}(g);\Sigma)+\frac{\widehat{\beta}}{K}\sum_{j\in\mathcal{N}_1}D(\boldsymbol{x}_j,\boldsymbol{\mu}(g);\Sigma)\right\}. \tag{6}$$

Estimates of label y_1 in formulas (5) and (6) are the same except for $\boldsymbol{\mu}(y_j)$ and $\boldsymbol{x}_j$ in the respective last terms. Note that $\boldsymbol{x}_j$ itself is a primitive estimate of $\boldsymbol{\mu}(y_j)$. Hence, the classification method based on the divergence model can be regarded as a natural extension of Switzer's method.

Switzer's method is known to give fairly good classification results in many practical situations. The local continuity assumption of the categories, however, cannot be applied to the whole image, and the procedure of estimating the parameters is not established. Thus, the divergence model can be seen as an extension of Switzer's method into the established MRF-based framework.

3.2. Error rates due to divergence model and Switzer's model

We will derive the exact error rate due to the divergence model and Switzer's model in the previous local region with two categories when $\mathcal{G} = \{1, 2\}$. In the two-category case, the only positive quasi-distance is $J(1, 2)$. Hence, by replacing $\beta J(1, 2)$ for β, we note that the MRF based on Jeffreys divergence is reduced to the Potts model.

Let δ be the Mahalanobis distance between distributions $N_{\mathrm{m}}(\boldsymbol{\mu}(g), \Sigma)$ for $g = 1, 2$, and $\mathcal{N}_1$ be a neighborhood consisting of $2K$ neighbors of pixel 1, where K is a fixed natural number. Furthermore, suppose that a number of neighbors with label 1 or 2 is randomly changing. Our aim is to derive the error rate of center pixel 1 given features $\boldsymbol{x}_1, \boldsymbol{x}_j$, and labels y_j of neighbors j in $\mathcal{N}_1$. Recall that $\widehat{Y}_{\mathrm{Div}}$ is the estimated label of y_1 obtained by formula (5). Then, exact error rate $\Pr\{\widehat{Y}_{\mathrm{Div}} \neq Y_1\}$ is given by

$$e(\widehat{\beta}; \beta, \delta) = \pi_0 \Phi(-\delta/2) + \sum_{k=1}^{K} \pi_k \left\{ \frac{\Phi(-\delta/2 - k\widehat{\beta}\delta/K)}{1 + e^{-k\beta\delta^2/K}} + \frac{\Phi(-\delta/2 + k\widehat{\beta}\delta/K)}{1 + e^{k\beta\delta^2/K}} \right\}, \tag{7}$$

where $\Phi(x)$ is the cumulative standard normal distribution function, and $\widehat{\beta}$ is an estimate of clustering parameter β. Here, π_k gives a prior probability such that the number of neighbors with label 1, L_1, is equal to $K + k$ or $K - k$ in set $\mathcal{N}_1$ for $k = 0, 1, \ldots, K$. Fig. 2 gives $L_1 = 2$ and $L_1 = 3$ in neighborhoods $\{2, 4, 6, 8\}$ and $\{2, 3, \ldots, 9\}$, respectively. See Appendix A for the derivation of formula (7).

If prior probability π_0 is equal to one, K pixels in neighborhood $\mathcal{N}_1$ are labeled 1 and the remaining K pixels are labeled 2 with probability one. In this case, the majority vote of neighbors does not work. Hence, we assume that π_0 is less than one. Then, we have the following properties of error rate $e(\widehat{\beta}; \beta, \delta)$, and they are shown in Appendix A.

P1. $e(0; \beta, \delta) = \Phi(-\delta/2)$, $\lim_{\widehat{\beta}\to\infty} e(\widehat{\beta}; \beta, \delta) = \pi_0 \Phi(-\delta/2) + \sum_{k=1}^{K} \dfrac{\pi_k}{1 + e^{k\beta\delta^2/K}}$.

P2. Function $e(\widehat{\beta}; \beta, \delta)$ of $\widehat{\beta}$ is minimized at $\widehat{\beta} = \beta$ (Bayes' rule), and minimum value $e(\beta; \beta, \delta)$ is a monotonically decreasing function of Mahalanobis distance δ for any fixed positive clustering parameter β.

P3. Function $e(\widehat{\beta}; \beta, \delta)$ is a monotonically decreasing function of β for any fixed positive constants $\widehat{\beta}$ and δ.

P4. We have the inequality: $e(\widehat{\beta}; \beta, \delta) < \Phi(-\delta/2)$ for any positive $\widehat{\beta}$ if $\beta \geqslant \dfrac{1}{\delta^2} \log\left\{ \dfrac{1 - \Phi(-\delta/2)}{\Phi(-\delta/2)} \right\}$ holds.

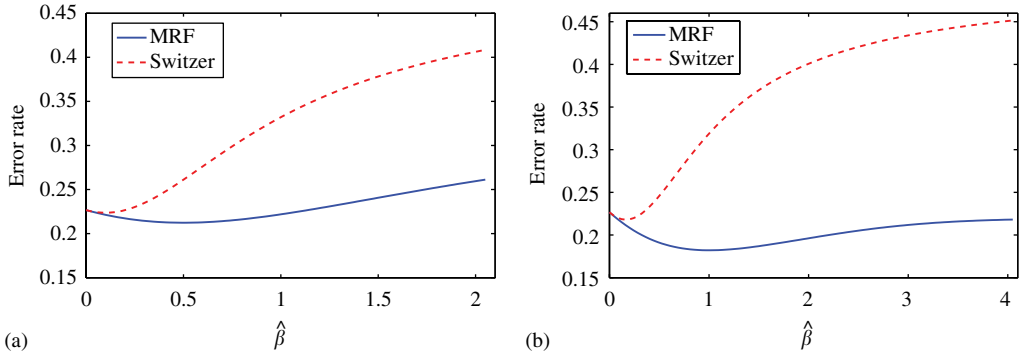

Fig. 3. Error rates due to MRF and Switzer's method: (a) $\delta = 1.5$, $\beta = 0.5$; (b) $\delta = 1.5$, $\beta = 1$.

Note that the value $e(0; \beta, \delta) = \Phi(-\delta/2)$ is simply the error rate due to Fisher's linear discriminant function with uniform priors on the labels. The asymptotic value $\sum_{k=1}^{K} \pi_k / \left(1 + e^{k\beta\delta^2/K}\right)$ given in P1 is the error rate due to the vote-for-majority rule performed by the neighbors when the number of neighbors, L_1, is not equal to K. Property P2 recommends us to use the true parameter β if it is known, and this is quite natural. Property P3 means that the classification becomes more efficient when δ and/or β becomes large. Note that δ is a distance in the feature space, and β is a distance in the image. Property P4 implies that the use of spatial information *always improves noncontextual discrimination*, even if estimate $\widehat{\beta}$ is far from true value β.

The error rate due to Switzer's method is obtained in the same form of as that of formula (7) by replacing δ with $\delta_* \equiv \delta \Big/ \sqrt{1 + 4\widehat{\beta}^2/K}$ $(< \delta)$ appearing in $\Phi(\cdot)$.

The comparison of these two error rates is illustrated in Fig. 3 with 4 $(2K)$ neighbors. Error rates due to the rules of formulas (5) and (6) against estimated clustering parameter $\widehat{\beta}$ for cases (a) $\delta = 1.5$ and $\beta = 0.5$, and (b) $\delta = 1.5$ and $\beta = 1$. The prior probability of random variable L_1 is defined by a binomial distribution with $\Pr\{L_1 = 2 \pm k\} = \binom{2}{k} \Big/ 4$ for $k = 0, 1, 2$.

The divergence model is seen to overcome Switzer's method, and both y-intercepts are the same value $\Phi(-\frac{1.5}{2}) = 0.2266$. Furthermore, the error rate due to the MRF is minimized at true value $\beta = 0.5$ or 1 (recall property P2). Parameters yielding Fig. 3(a) do not meet the sufficient condition of P4 because $\beta = 0.5 < 0.5455 = \log\left[\frac{1-\Phi(-\delta/2)}{\Phi(-\delta/2)}\right]/\delta^2$. Actually, the error rate exceeds the y-intercept $\Phi(-\delta/2)$ for large $\widehat{\beta}$, see Fig. 3(a). Whereas, those of (b) meet the condition because $\beta = 1$ in case (b) and $\delta = 1.5$ is common. Hence, the error rate is always less than $\Phi(-\delta/2)$ for any $\widehat{\beta}$, and this is confirmed by observing Fig. 3(b).

4. Parameter estimation in MRFs based on divergence

We consider the parameter estimation procedure for the divergence model formulated in Section 2. Suppose that a set of training data $\{(\boldsymbol{x}_i, y_i) \in \mathbb{R}^m \times \mathcal{G} \mid i \in \mathcal{D}\}$ is given. Set $\boldsymbol{x} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_n^T)^T$ and $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, and let $\Theta = \{\boldsymbol{\theta}(1), \ldots, \boldsymbol{\theta}(G)\}$ be a set of unknown parameters specifying category-specific densities of formula (1). We define the log conditional likelihood (1) of $\boldsymbol{x}|\boldsymbol{y}$ and

Table 1
Error rates (%) of classification results based on Gaussian MRFs with two sorts of divergences for the data grss_dfc_0006

Neighborhoods		Gaussian MRFs			
Radius	Average	Homoscedastic		Heteroscedastic	
r	of $\lvert\mathcal{N}_i\rvert$	Potts	Jeffreys	Potts	Jeffreys
0	0.00	8.61	8.61	14.67	14.67
2	3.46	6.44	**6.09**	12.93	**11.44**
3	6.73	6.22	**5.69**	13.09	**10.19**
4	9.72	5.97	**5.35**	12.80	**9.67**
5	15.59	6.35	**5.69**	13.16	**9.39**
6	21.02	6.49	**5.64**	13.16	**9.55**
7	26.27	6.68	**6.13**	12.90	**9.90**
8	33.72	6.70	**6.65**	12.92	**9.76**

Numerals in bold face imply the superiority of the divergence model.

the log pseudo-likelihood of $\boldsymbol{y}$ as

$$\ell_1(\Theta) = \sum_{i\in\mathcal{D}} \log f(\boldsymbol{x}_i, \boldsymbol{\theta}(y_i)) \quad \text{and} \quad \ell_2(\Theta, \beta) = \sum_{i\in\mathcal{D}} \log p_i(y_i \mid \boldsymbol{y}_{-i}), \tag{8}$$

respectively. Parameter set Θ appears in both likelihoods. Therefore, this causes difficulties in parameter estimation. Now, we propose the following three methods for parameter estimation.

The first estimation method is obtained by using the training and the test data sets. Parameter set Θ is estimated by using the training data as $\widehat{\Theta} = \arg\max_\Theta \ell_1(\Theta)$. Then, clustering parameter β as well as labels of test data are estimated by maximizing log pseudo-likelihood $\ell_2(\widehat{\Theta}, \beta)$ *of the test data*. This method is the simplest approach, and it could be widely used at various settings.

The second method is derived by only using the training data. Using estimate $\widehat{\Theta}$ from the previous method, we maximize the pseudo-likelihood *of the training data* and define $\widehat{\beta}(\widehat{\Theta}) = \arg\max_\beta \ell_2(\widehat{\Theta}, \beta)$. The estimating equation for β is given as follows:

$$\partial\ell_2(\widehat{\Theta}, \beta)/\partial\beta = \sum_{i\in\mathcal{D}} \sum_{g\in\mathcal{G}} \Delta_i(g) \left\{\delta_{g,y_i} - p_i(g \mid \boldsymbol{y}_{-i})\right\} = 0, \tag{9}$$

where $\Delta_i(g)$ is defined by formula (3). We will use a gradient algorithm for obtaining the optimal value $\widehat{\beta}(\widehat{\Theta})$, and that would be a feasible task.

The third method also only uses the training data. The optimal parameter set $\widehat{\Theta}(\beta) = \arg\max_\Theta \{\ell_1(\Theta) + \ell_2(\Theta, \beta)\}$ is found for fixed β. Then, the optimal value of β is chosen by $\arg\max_\beta\{\ell_1(\widehat{\Theta}(\beta)) + \ell_2(\widehat{\Theta}(\beta), \beta)\}$. This is repeated until the convergence condition is met. An iterative method is developed in Appendices B.1 and B.2 in the normal case as well as in a general exponential family case in B.3.

5. Applications to actual data

Our method was applied to benchmark data set grss_dfc_0006 provided by the IEEE Geoscience and Remote Sensing Society Data Fusion reference database [6]. The data consist of samples with fifteen variables ($m = 15$) and five agricultural categories ($G = 5$) observed at Feltwell in the U.K. The training and the test areas consist of 5072 and 5760 pixels, respectively.

Neighborhood $\mathcal{N}_i$ of pixel i is defined by a set of pixels whose distances from i are not greater than given radius r. If $r = 1, \mathcal{N}_i$ is given by the first-order neighborhood. We applied GMRFs with the following four possible combinations: homoscedastic or heteroscedastic case; and the Potts model or the divergence model for each case. The parameters are estimated by the first method described in Section 4. We classify the test data based on clustering parameter $\beta = 0(.05)10.00$ for fixed radius r. Then, the optimal β is chosen by maximizing the log pseudo-likelihood of the form ℓ_2 in formula (8).

Error rates of the optimal results with radius r in set $\{0, 2, 3, \ldots, 8\}$ are listed in Table 1. Both spatial models are seen to significantly improve the noncontextual classification result with $r = 0$. In the homoscedastic case, the divergence model exhibits a similar performance to that of the Potts model. In the heteroscedastic case, the divergence model is superior to the Potts model.

6. Conclusion

We have considered the MRF specified by Jeffreys divergence between category-specific densities with emphasis on normal distributions. Furthermore, the divergence model is compared with Switzer's smoothing method. Parameter estimation methods are also established.

We summarize the features of the divergence model as follows.

- The classification based on the divergence model is a natural extension of Switzer's smoothing method.
- The error rate of our method was obtained in the closed form under the simple setup. The use of spatial information was shown to always reduce the error rate under certain condition.
- The divergence model was applied to the actual benchmark data set. Then, the proposed model exhibits a better performance than that of the Potts model.

The divergence model can be defined in an exponential family of probability distribution. The parameter estimation method of the general family of probability densities is developed in Appendix B.3.

We have proposed three estimation procedures in Section 4. However, the problem of determining which estimation method is efficient for given data still remains. In addition, model selection is another important problem.

Acknowledgments

The authors are grateful for the reviewers and the associate editor for their constructive comments. Data set grss_dfc_0006 is provided by the IEEE GRS-S Data Fusion committee. This research was supported by a Grant-in-Aid for Scientific Research (C) 15540123 from the Japan Society for the Promotion of Science.

Appendix A. Error rates in local region

Let L_1 be a random variable denoting a number of neighbors with label 1 in $\mathcal{N}_1$. Then, the remaining $2K - L_1$ neighbors are labeled 2 for $L_1 = 0, 1, \ldots, 2K$. The label of the center pixel is estimated by formula (5). According to formula (2), $\Pr\{Y_1 = 1 \mid L_1 = K + k\} = 1/\{1 + \exp(-k\beta\delta^2/K)\}$ and $\Pr\{Y_1 = 2 \mid L_1 = K + k\} = 1/\{1 + \exp(k\beta\delta^2/K)\}$ hold

for $k = 0, \pm 1, \ldots, \pm K$. Due to the classification rule of formula (5), the following can be obtained.

$$\Pr\{\widehat{Y}_{\mathrm{Div}} \neq Y_1 \mid Y_1 = 1, L_1 = K + k\} = \Phi\left(-\delta/2 - k\widehat{\beta}\delta/K\right),$$

$$\Pr\{\widehat{Y}_{\mathrm{Div}} \neq Y_1 \mid Y_1 = 2, L_1 = K + k\} = \Phi\left(-\delta/2 + k\widehat{\beta}\delta/K\right)$$

for $k = 0, \pm 1, \ldots, \pm K$. Thus, we have the relationship: $\Pr\{\widehat{Y}_{\mathrm{Div}} \neq Y_1 \mid L_1 = K + k\} = \Pr\{\widehat{Y}_{\mathrm{Div}} \neq Y_1 \mid L_1 = K - k\}$. Its value, $e_k(\widehat{\beta}; \beta, \delta)$, is given by

$$e_k(\widehat{\beta}; \beta, \delta) = \frac{\Phi(-\delta/2 - k\widehat{\beta}\delta/K)}{1 + e^{-k\beta\delta^2/K}} + \frac{\Phi(-\delta/2 + k\widehat{\beta}\delta/K)}{1 + e^{k\beta\delta^2/K}}. \tag{A.1}$$

Taking the expectation with respect to L_1, we have the error rate presented in formula (7), where $\pi_k = \Pr\{L_1 = K \pm k\}$ is the prior distribution of L_1.

Property P1 follows immediately. Property P4 is obtained by the sufficient condition such that $e(0; \beta, \delta) > \lim_{\widehat{\beta} \to \infty} e(\widehat{\beta}; \beta, \delta)$. Now, we only need to show that the conditional error rate (A.1) satisfies P2 and P3 because the error rate presented in formula (7) is given by their convex combination. This is shown as follows.

Set $k' = k/K$, $\xi = 1/\{1 + \exp(k'\beta\delta^2)\}$, and $\phi(x) = \exp(-x^2/2)/\sqrt{2\pi}$. Noting the relationship $\phi(-\delta/2 - k'\beta\delta)e^{k'\beta\delta^2} = \phi(-\delta/2 + k'\beta\delta)$, we have the following derivatives:

$$\partial e_k(\widehat{\beta}; \beta, \delta)/\partial\widehat{\beta} = k'\delta\xi\phi(-\delta/2 + k'\beta\delta)\{e^{k'\delta^2(\widehat{\beta} - \beta)} - 1\}, \tag{A.2}$$

$$\begin{aligned}\partial e_k(\beta; \beta, \delta)/\partial\delta = &-\xi\phi(-\delta/2 + k'\beta\delta) \\ &- 2k'\beta\delta\xi^2 e^{k'\beta\delta^2}\{\Phi(-\delta/2 + k'\beta\delta) - \Phi(-\delta/2 - k'\beta\delta)\},\end{aligned} \tag{A.3}$$

$$\partial e_k(\widehat{\beta}; \beta, \delta)/\partial\beta = -k'\delta^2\xi^2 e^{k'\beta\delta^2}\{\Phi(-\delta/2 + k'\widehat{\beta}\delta) - \Phi(-\delta/2 - k'\widehat{\beta}\delta)\}. \tag{A.4}$$

Derivative (A.2) implies that function $e_k(\widehat{\beta}; \beta, \delta)$ of $\widehat{\beta}$ is minimized at β. Derivative (A.3) implies that minimum value $e_k(\beta; \beta, \delta)$ is a monotonically decreasing function of δ for any positive β because of the monotonicity of the function $\Phi(\cdot)$. Similarly, derivative (A.4) yields P3.

Appendix B. Parameter estimation based on sum of log likelihoods

Let $\ell(\Theta, \beta) = \ell_1(\Theta) + \ell_2(\Theta, \beta)$ be the sum of the log likelihoods defined by formula (8). Then, the estimating equations for the optimal parameter $\widehat{\Theta}_\beta = \arg\max_\Theta\{\ell(\Theta, \beta)\}$ with fixed β will be obtained in the following three cases.

B.1. Homoscedastic GMRFs

Consider normal distributions $N_m(\boldsymbol{\mu}(g), \Sigma)$ with common variance–covariance matrix Σ for label g. In this case, the sum of the log likelihoods is given by

$$\begin{aligned}\ell = &-mn\ \log(2\pi)/2 - n\ \log|\Sigma|/2 \\ &- \sum_{i \in \mathcal{D}} \left[\{\boldsymbol{x}_i - \boldsymbol{\mu}(y_i)\}^T \Sigma^{-1}\{\boldsymbol{x}_i - \boldsymbol{\mu}(y_i)\}/2 + \beta\Delta_i(y_i)/2 + \log\left\{\sum_{h \in \mathcal{G}} \exp\left(-\beta\Delta_i(h)\right)\right\}\right],\end{aligned}$$

where $\Delta_i(\cdot)$ is defined by formula (3) with $J(g,h) = D(\boldsymbol{\mu}(g), \boldsymbol{\mu}(h); \Sigma)$: the squared Mahalanobis distance. We define the following notations:

$$\bar{r}_g(h) = \sum\nolimits_{i\in\mathcal{D}(g)} r_i(h), \quad \bar{p}_+(h) = \sum\nolimits_{i\in\mathcal{D}} p_i(h \mid \boldsymbol{y}_{-i}),$$
$$\bar{r}_+(h) = \sum\nolimits_{i\in\mathcal{D}} r_i(h) \quad \text{and} \quad \bar{d}_+(g,h) = \sum\nolimits_{i\in\mathcal{D}} r_i(g)\, p_i(h \mid \boldsymbol{y}_{-i}) \tag{B.1}$$

for $g, h \in \mathcal{G}$, where $\mathcal{D}(g)$ is a set of pixels with label g in training area $\mathcal{D}$, and $p_i(\cdot|\cdot)$ and $r_i(\cdot)$ are, respectively, defined by formulas (2) and (3). The estimating equations for mean vectors $\boldsymbol{\mu}(g)$ and covariance matrix Σ can then be obtained by differential equations $\partial\ell/\partial\boldsymbol{\mu}(g) = \mathbf{0}$ and $\partial\ell/\partial\Sigma = O$ as

$$n_g\,[\bar{\boldsymbol{x}}(g) - \boldsymbol{\mu}(g)] - 2\beta \sum_{h\in\mathcal{G}} a(g,h)\{\boldsymbol{\mu}(g) - \boldsymbol{\mu}(h)\} = \mathbf{0}, \tag{B.2}$$

and

$$\Sigma - \frac{1}{n}\sum_{g\in\mathcal{G}}\sum_{i\in\mathcal{D}(g)} \{\boldsymbol{x}_i - \boldsymbol{\mu}(g)\}\{\boldsymbol{x}_i - \boldsymbol{\mu}(g)\}^T - \beta T = O, \tag{B.3}$$

where $\bar{\boldsymbol{x}}(g) = \sum_{i\in\mathcal{D}(g)} \boldsymbol{x}_i/n_g$, $n_g = |\mathcal{D}(g)|$,

$$T = \frac{2}{n}\sum_{h\in\mathcal{G}}\sum_{h'\in\mathcal{G}} b(h,h')\{\boldsymbol{\mu}(h) - \boldsymbol{\mu}(h')\}\{\boldsymbol{\mu}(h) - \boldsymbol{\mu}(h')\}^T, \tag{B.4}$$

and

$$a(g,h) = b(g,h) + b(h,g), \quad b(g,h) = \bar{r}_g(h) - \bar{d}_+(h,g). \tag{B.5}$$

We note that $a(g,h)$ depends on the Mahalanobis distance, so unknown parameters $\boldsymbol{\mu}(g)$ and Σ should be estimated by an iterative procedure. Here, we recommend taking sample means and the sample variance–covariance matrix as initial estimates.

B.2. Heteroscedastic GMRFs

Second, we investigate the case of normal distributions $N_m(\boldsymbol{\mu}(g), \Sigma(g))$ with different variance–covariance matrices for g in $\mathcal{G}$. Then, Jeffreys divergence $J(g,h)$ is calculated as

$$\begin{aligned} J(g,h) = {} & \{\boldsymbol{\mu}(g) - \boldsymbol{\mu}(h)\}^T\{\Sigma(g)^{-1} + \Sigma(h)^{-1}\}\{\boldsymbol{\mu}(g) - \boldsymbol{\mu}(h)\}/2 \\ & + \operatorname{trace}\left\{\Sigma(h)^{-1}\Sigma(g) + \Sigma(g)^{-1}\Sigma(h)\right\}/2 - m. \end{aligned} \tag{B.6}$$

The estimating equations for $\boldsymbol{\mu}(g)$ and $\Sigma(g)$ with $g \in \mathcal{G}$ are expressed by

$$n_g\,[\bar{\boldsymbol{x}}(g) - \boldsymbol{\mu}(g)] - 2\beta \sum_{h\in\mathcal{G}} b(g,h)\{\boldsymbol{\mu}(g) - \boldsymbol{\mu}(h)\}$$

$$-\beta \sum_{h\in\mathcal{G}} b(h,g)\{E_{\mathrm{m}} + \Sigma(g)\Sigma(h)^{-1}\}\{\boldsymbol{\mu}(g) - \boldsymbol{\mu}(h)\} = \mathbf{0},$$

$$n_g\Sigma(g) - \sum_{i\in\mathcal{D}(g)} \{\boldsymbol{x}_i - \boldsymbol{\mu}(g)\}\{\boldsymbol{x}_i - \boldsymbol{\mu}(g)\}^T + \Sigma(h) - \Sigma(g)\Sigma^{-1}(h)\Sigma(g)$$

$$-\beta \sum_{h\in\mathcal{G}} a(g,h)\{\boldsymbol{\mu}(g)-\boldsymbol{\mu}(h)\}\{\boldsymbol{\mu}(g)-\boldsymbol{\mu}(h)\}^T = O,$$

where E_{m} stands for the identity matrix, and $a(g,h)$ and $b(g,h)$ are defined in (B.5).

B.3. Extension to natural exponential family and Jeffreys divergence

We proceed to the general case in which category-specific densities are of the exponential family. Let the densities in formula (1) belong to an exponential family of the form

$$f(\boldsymbol{x}_*,\boldsymbol{\theta}) = f_0(\boldsymbol{x}_*)\exp\{\boldsymbol{t}(\boldsymbol{x}_*)^T\boldsymbol{\theta}-\kappa(\boldsymbol{\theta})\} \tag{B.7}$$

for $\boldsymbol{\theta}=\boldsymbol{\theta}(g)$ with $g\in\mathcal{G}$, where $\boldsymbol{t}(\boldsymbol{x}_*)$ is a vector of sufficient statistics of m-dimensional feature vector $\boldsymbol{x}_*$, and $\kappa(\boldsymbol{\theta})$ is the cumulant transform. Then, in this case Jeffreys divergence $J(g,h)$ is given by

$$J(g,h) = \{\boldsymbol{\eta}(g)-\boldsymbol{\eta}(h)\}^T\{\boldsymbol{\theta}(g)-\boldsymbol{\theta}(h)\}, \tag{B.8}$$

where $\boldsymbol{\eta}(g)=\partial\kappa(\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(g)}$ is a mean vector.

We will derive an iterative algorithm for obtaining the pseudo maximum likelihood estimate: $\widehat{\Theta}_\beta = \arg\max_\Theta\{l_1(\Theta)+l_2(\Theta,\beta)\}$ for given $\beta\geqslant 0$, where $l_1(\Theta)$ and $l_2(\Theta,\beta)$ are the log likelihoods derived by density (B.7) and divergence (B.8). First, we prepare the following relationship due to partial derivatives for $\boldsymbol{\theta}(g)$, as

$$\frac{\partial J(h,h')}{\partial\boldsymbol{\theta}(g)} = \{\delta_{gh}I(h)-\delta_{gh'}I(h')\}\{\boldsymbol{\theta}(h)-\boldsymbol{\theta}(h')\}+(\delta_{gh}-\delta_{gh'})\{\boldsymbol{\eta}(h)-\boldsymbol{\eta}(h')\},$$

where $I(g)\equiv\partial^2\kappa(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T|_{\boldsymbol{\theta}=\boldsymbol{\theta}(g)} : m\times m$ is the Fisher information. Second, using the above formula and relationships $\sum_{g\in\mathcal{G}} r_i(g)=\sum_{g\in\mathcal{G}} p_i(g|\boldsymbol{y}_{-i})=1$, we have

$$\frac{\partial}{\partial\boldsymbol{\theta}(g)}\sum_{i\in\mathcal{D}}\Delta_i(y_i) = \sum_{h\in\mathcal{G}}\{\bar{r}_g(h)+\bar{r}_h(g)\}\boldsymbol{\tau}(g,h),$$

and

$$\frac{\partial}{\partial\boldsymbol{\theta}(g)}\sum_{i\in\mathcal{D}}\log\left[\sum_{h\in\mathcal{D}}\exp\{-\beta\Delta_i(h)\}\right] = -\beta\sum_{i\in\mathcal{D}}\sum_{h\in\mathcal{G}}\{\bar{d}_+(g,h)+\bar{d}_+(h,g)\}\boldsymbol{\tau}(g,h),$$

where $\boldsymbol{\tau}(g,h)=I(g)\{\boldsymbol{\theta}(g)-\boldsymbol{\theta}(h)\}+\boldsymbol{\eta}(g)-\boldsymbol{\eta}(h)$, and $\bar{d}_+(g,h)$ is defined by (B.1) with $J(g,h)$ in formula (B.8).

Finally, the estimating equation $\partial\{\ell_1(\Theta)+\ell_2(\Theta,\beta)\}/\partial\boldsymbol{\theta}(g)=\mathbf{0}$ is given by

$$n_g\{\bar{\boldsymbol{t}}(g)-\boldsymbol{\eta}(g)\}-\beta\sum_{h\in\mathcal{G}}a(g,h)\boldsymbol{\tau}(g,h)=\mathbf{0},$$

where $a(g,h)$ is defined by (B.5) with $J(g,h)$ of (B.8).

References

[1] J. Besag, On the statistical analysis of dirty pictures, J. Roy. Statist. Soc. Ser. B 48 (1986) 259–302.
[2] J.P. Chilès, P. Delfiner, Geostatistics, Wiley, New York, 1999.

[3] N. Cressie, Statistics for Spatial Data, second ed., Wiley, New York, 1993.
[4] S. Eguchi, J. Copas, A class of logistic-type discriminant functions, Biometrika 89 (2002) 1–22.
[5] S. Geman, D. Geman, Stochastic relaxation, Gibbs distribution, and Bayesian restoration of images, IEEE Trans. Pattern Anal. Mach. Intell. 6 (1984) 721–741.
[6] IEEE GRSS Data Fusion reference database, 2001, The data set GRSS_DFC_0006. Online ⟨http://www.dfc-grss.org/⟩.
[7] A.K. Jain, R.P.W. Duin, J. Mao, Statistical pattern recognition: a review, IEEE Trans. Pattern Anal. Mach. Intell. 22 (2000) 4–37.
[8] H. Jeffreys, An invariant form for the prior probability in estimation problems, Proc. Roy. Soc. London, Ser. A 186 (1946) 453–461.
[9] S.Z. Li, Modeling image analysis problems using Markov random fields, Handbook of Statistics, vol. 20, Wiley, New York, 2000, pp. 1–43.
[10] K.V. Mardia, Multi-dimensional multivariate Gaussian Markov random fields with application to image processing, J. Multivariate Anal. 24 (1988) 265–284.
[11] J. Marroquin, F.A. Velasco, M. Rivera, M. Nakamura, Gauss–Markov measure field models for low-level vision, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2001) 337–348.
[12] G.J. McLachlan, Discriminant Analysis and Statistical Pattern Recognition, second ed., Wiley, New York, 2004.
[13] R. Nishii, A Markov random field-based approach to decision level fusion for remote sensing image classification, IEEE Trans. Geosci. Remote Sensing 41 (2003) 2316–2319.
[14] R. Nishii, S. Eguchi, Supervised image classification by contextual AdaBoost based on posteriors in neighborhoods, IEEE Trans. Geosci. Remote Sensing 43 (2005) 2547–2554.
[15] P. Switzer, Extensions of linear discriminant analysis for statistical classification of remotely sensed satellite imagery, Math. Geol. 12 (1980) 367–376.
[16] I.J. Taneja, New developments in generalized information measures, in: P.W. Hawkes (ed.), Advances in Imaging and Electron Physics, vol. 91, 1995, pp. 37–135.